基于 XGBoost 的单脉冲信号识别研究

凌毓 张金区* 李乡儒 李慧

华南师范大学计算机学院, 广东广州, 510631

摘要：脉冲星搜寻是射电天文学领域的重要研究方向，随着大型射电望远镜的不断建设和发展，采集的数据呈指数级增长，如何及时从快速获取的海量数据中准确识别出脉冲星信号成为当前面临的巨大挑战。本文以 LOFAR 联合阵列巡天项目的观测数据为例，设计了针对单脉冲信号识别的 10 个特征变量，进一步研究了 XGBoost 结合包裹式特征选择法在单脉冲信号识别中的应用，并对比分析了 GBDT、AdaBoost、随机森林和 BP 神经网络等模型对单脉冲信号识别的实验效果。实验结果表明，XGBoost 结合包裹式特征选择法在单脉冲信号识别方面更具综合优势，其误分类率最低，同时分类结果的精确率、召回率与 F1-score 值都最高，平均高出其它模型 1 到 2 个百分点。从特征选择上来说，有九个特征被选为最优特征。本研究设计的特征变量和识别方法可为我国开展以 FAST 探测信号为主的脉冲星搜寻提供方法和技术参考。

关键词：单脉冲；XGBoost；特征选择；包裹法

中图分类号：P162.5　　文献标识码：A

1. 引言

脉冲星是快速旋转的中子星，因不断地发出电磁脉冲信号而得名。脉冲星的发现是二十世纪六十年代的天文学重要发现之一[1]。对脉冲星的研究可以为拥有极端物理属性的中子星研究提供重要信息，能极大地推动物理、天文、导航和时间度量等领域的发展[2][3]。基于脉冲星信号的周期性特征，利用快速傅里叶变换将时间域信号转换为频率域信号，并结合快速折叠算法成为探测脉冲星信号的主要方法[4]。随着对脉冲信号的挖掘，近年来相继发现了两类没有周期性特征的天文脉冲信号，它们分别来自旋转瞬间射电体(Rotating Radio Transients, RRATs)和快速射电暴(Fast Radio Bursts, FRBs)[2][5][6]。旋转瞬间射电体信号的发射在时间上非常零星和分散，无法在传统的周期性搜寻中发现它们。快速射电暴由银河系外的射电突发信号组成，目前发现极个别的周期性现象，但仍然以缺乏周期性为主要特征。这两类天文现象的脉冲信号，由于它们转瞬即逝的特征，被称之为单脉冲天体信号。单脉冲信号的搜寻不但是周期性信号搜寻方法的有益补充，而且是旋转瞬间射电体和快速射电暴的主要探测方法[4]。

目前，单脉冲信号的识别方法主要分为启发式阈值搜寻算法和机器学习算法。启发式搜

*基金项目：国家自然科学基金（11973022, 61273248, 61075033）
作者简介：凌毓，女，硕士研究生，研究方向：脉冲信号处理，邮箱：2019022610@m.scnu.edu.cn
通讯作者：张金区，男，博士，副教授，研究方向：空间数据处理，邮箱：zjq@scnu.edu.cn

寻是一种利用问题所具有的启发信息来引导搜寻、发现目标的算法，该算法通过减小搜寻范围来降低问题复杂性，提升计算效率。这类方法主要是基于 Cordes 和 McLaughlin 提出的单脉冲信号分类框架[2]。该框架将单脉冲信号的提取分为消色散、匹配滤波、阈值化和判断等四个步骤，以确定被检测信号中单脉冲信号的存在性。例如，Deneva 等人使用一种聚类算法将信噪比高于一定阈值的疑似脉冲事件筛选为单脉冲候选体[7]。Karako Argaman 等人根据色散（DM）和信号时间对脉冲事件进行分组，然后基于相邻分组中最大信噪比判断脉冲信号是否有峰值发生，据此筛选并创建诊断图以供人工检查[8]。Ryan 等人进一步提出了一种简单的递归峰值识别算法，利用弥散脉冲组（DPG）拟合线的斜率来识别 DPG 的大斜率趋势，并据此判断单脉冲事件候选体[9]。这些方法虽然在检测脉冲信号时有一定的作用，但是它们主要依靠阈值分割来提取脉冲星信号，所采用的特征来源于分组中最强的脉冲信号，由此导致它们往往精度有限，且需要大量人工参与，难以适应大规模、海量的数据处理。

近年来，随着传感器技术的发展和大规模射电巡天的推进，机器学习已经成为脉冲星信号识别的重要途经[10]。机器学习方法是通过对已知脉冲星信号的特征进行统计分析，建立学习模型，然后利用学习模型对未知脉冲信号进行判断的方法。该方法通常需要四个步骤：（1）建立基准数据集；（2）特征提取；（3）模型训练与评价；（4）模型应用。McFadden 等人在总结机器学习在脉冲信号筛选中的应用时，指出目前已有的机器学习算法主要用于周期性脉冲信号的搜寻[11]。例如，人工神经网络（ANN）算法[12-14]和模式识别算法[15]都在周期性脉冲信号搜寻中进行了应用。虽然机器学习在周期性脉冲信号中已经有了许多探索，但在单脉冲信号识别中的应用才刚刚开始并逐渐受到重视。在单脉冲信号识别的机器学习应用方面，Eatough 等在启发式阈值搜寻算法的基础上，挑选了信噪比、脉冲宽度等 12 个特征作为三层人工神经网络的输入，首次以机器学习的方法进行单脉冲信号筛选[12]。Ryan 等人利用 Green Bank 望远镜观测到的数据集，从脉冲数量-色散图、信噪比-色散图中提取了 16 个特征，比较了 SVM、ANN、RULE 和决策树等方法，实验结果指出使用随机森林集成树的分类器在查全率和查准率方面提供了最佳的整体效果[9]。Michilli 等人以 LOTAAS 数据集为例，根据每个特征的信息增益，筛选了用于单脉冲信号分类的五个重要指标：峰值检测窗口宽度、脉冲色散平均值、脉冲信噪比、窗口宽度分布曲线超额峰度、以及信噪比分布曲线超额峰度（关于各个指标的进一步解释请见本文 2.3 节）[16]。该工作通过比较几种不同的机器学习算法，认为基于高斯-海灵格快速决策树的方法在单脉冲信号分类中具有最好的性能。

从以往的研究来看，基于决策树的方法被认为是性能最好的方法之一，但是对决策树模型的参数估计大多采用了小规模随机抽样的方法来计算，无法保证最终分类结果的最优性。近些年机器学习领域又对决策树模型进行了改进和提升，尤其是基于梯度提升的 GBDT 和 XGBoost 算法在许多领域都得到了广泛的应用[17]。因此，本文旨在探讨 XGBoost 结合包裹式特征选择方法进行单脉冲信号识别的性能分析。接下来，论文第二部分介绍用于研究的数据集，第三部分详细说明 XGBoost 算法原理，第四部分为实验结果与对比分析讨论，最后

进行了总结和展望。

2. 数据

2.1 数据来源

良好的基准数据集是进行机器学习训练应用和研究的基础，然而面对海量的脉冲信号，对脉冲信号进行标注是短期难以完成的事情。因此在本文中，我们直接使用 Michilli 工作中已标注的数据集用于模型的研究[15]。此数据集来源于低频射电联合阵列巡天（LOFAR tied-array all-sky survey，LOTAAS）项目。低频射电阵列（Low frequency array，LOFAR）由荷兰射电天文研究所带头主持研发，是一种由数千个天线组成的大型射电望远镜，这些天线被分组分布在荷兰和其他欧洲国家的观测站点中。在最低频段以高分辨率和高灵敏度用于进行脉冲星广泛且深入的研究[18]。LOTAAS 项目则利用了其中 12 个子站进行观测，对天空中的每个指向产生 222 个同时段的射电数据，每次观测时间持续 1 个小时，记录数据的时间分辨率为 0.492 毫秒，每小时可以接受 16.9TB 的原始数据[19]。本文实验所用数据集是从历次 LOTAAS 观测中抽取的。

2.2 数据预处理过程

脉冲星的搜寻大致需要四个阶段，分别是射电信号数据收集、消色散处理、周期性脉冲或单脉冲搜寻，以及人工判别[3]，其中色散效应是天体物理信号和 RFI 信号的重要区别之一[20]。天体物理信号到达地球时受空间中不同密度自由电子的影响，导致不同频率的信号产生不同的延迟效应。色散（DM）是对信号传播方向上自由电子总数量的度量。由于事先并不知道天体信号对应的 DM，所以在做消色散处理时，需要用不同的 DM 值来尝试。由此可知，对于一个单脉冲信号而言，虽然其本质上对应着唯一一个 DM，但经过消色散的处理过程，会生成很多根据不同 DM 消色散得到的候选脉冲信号，这些不同 DM 对应的候选脉冲信号，仍然可能被检测为峰值信号。这样，理论上的一个脉冲信号可能会被检测为多个峰值信号，它们对应的 DM 值非常接近。因此，可通过对被检测到的一系列峰值信号按照对应的 DM 值进行聚类分析，聚集到一起的峰值信号形成一个弥散脉冲组(Dispersed Pulse Groups, DPG)。图 1 中第 1 个子图展示了一个弥散脉冲组中不同 DM 值下得到的脉冲信号的信噪比分布。对单脉冲信号的识别，主要是识别弥散脉冲组（DPG）是来源于脉冲星还是 RFI，如果被识别为脉冲星信号，则进一步输出特征图信息供人工进一步判断。

本文所采用的数据，是在 0 到 550 pc cm^{-3} 的色散范围内，对 DM 每间隔 0.01 到 0.1 pc cm^{-3} 进行一次计算处理后得到的。对 DM 处理后的数据，采用不同长度的矩形窗进行峰值检测，将信噪比大于 5 的信号进行保存，形成一个信号事件表，保存的信息包括窗口宽度、色散、信号时间等。基于信号事件表中每条记录的信号时间和 DM 值的邻近程度，对信号事件进行聚类分组，对信号时间在 30 毫秒内、DM 差值在 2 pc cm^{-3} 范围内的信号事件归为一个弥散脉冲组（DPG）。图 1 展示了脉冲星编号为 B1133+16 的一个弥散脉冲组的信号事件分布情况，以及一个射频干扰信号构成的弥散脉冲组的分布情况。从中可以看出，脉冲星信号和

RFI 信号的弥散脉冲组在信噪比分布曲线形态上有显著差异，其窗口宽度的分布曲线也有明显差异，这些形态特征有助于脉冲星弥散脉冲组的识别。

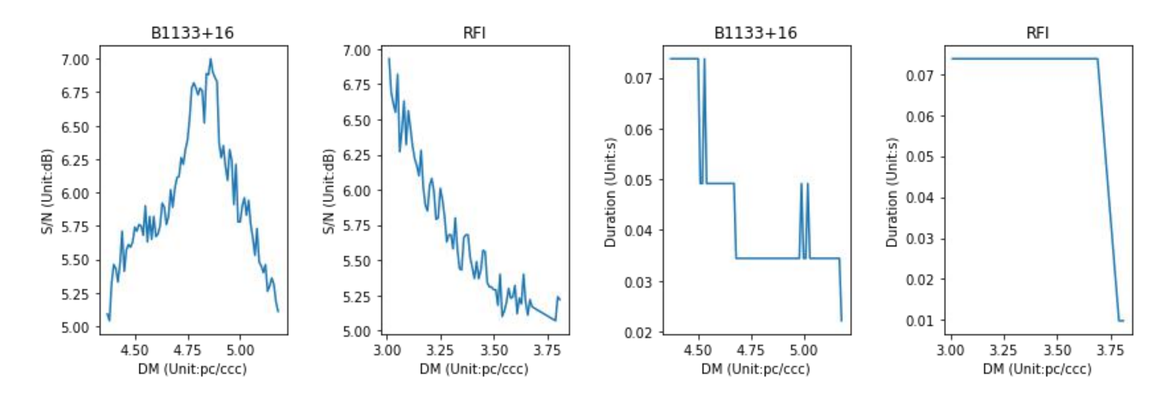


图 1 单脉冲和 RFI 弥散脉冲组（DPG）事件分布示意图

Fig. 1 DPG distribution curve of S/N and Duration for single pulse and RFI

2.3 数据特征设计

经过滤和峰值检测筛选后，信号事件表记录的总记录数约 374 万条，形成的弥散脉冲组（DPG）53066 条，其中 35063 条为射频干扰（RFI）记录，18003 条属于 47 个已知脉冲星的脉冲记录。对弥散脉冲组（DPG）的特征设计是进行正确分类的重要内容，参考已有弥散脉冲组（DPG）特征应用的方法，本文设计以下特征：

（1）色散（DM），是脉冲星和地球之间沿信号传播方向的自由电子积分柱密度，单位 pc cm^{-3}，一个弥散脉冲组（DPG）的 DM 值取其中最强的信号事件对应的 DM 值。

（2）信噪比（S/N），是信号和噪声的比值，即射电望远镜接收到信号的电压值与同时记录的噪声电压的比值。信噪比越高，即信号越强，噪声越弱。信噪比是判断脉冲事件的主要依据，一个弥散脉冲组（DPG）的 S/N 取其中最强的信号事件对应的 S/N 值。

（3）窗口宽度（Duration），对时间序列信号进行峰值检测时，所用的矩形窗函数的窗口宽度，即窗口的时间范围，是用于峰值提取的计算参数。在进行峰值检测时，采用了一系列不同的窗口宽度进行检测，不同的窗口宽度可能检测出不一样的峰值结果。一个弥散脉冲组（DPG）的窗口宽度取其中最强的信号事件对应的窗口宽度。

（4）色散范围（DM Extent,缩写为 DM_E），是一个弥散脉冲组（DPG）中所有信号事件对应的 DM 值范围，即图 1 中第 1 个子图中曲线的覆盖范围。

（5）时间范围（Time Extent,缩写为 Time_E），是一个弥散脉冲组（DPG）中所有信号事件对应的时间范围，单位秒。

（6）事件个数（Number of Events,缩写为 N_Events），是一个弥散脉冲组（DPG）包含的信号事件的数量，事件数量太少则说明没有太强的色散效应，大概率不是脉冲星信号。

（7）色散平均值（Average DM,缩写为 aDM），属于同一个弥散脉冲组（DPG）的所有信号事件的平均色散。

（8）脉冲平均时间（Average Time of Pulse，缩写为 aTime），形成一个弥散脉冲组（DPG）的所有信号的平均时间。因为 LOTAAS 项目利用了 12 个子站同时进行观测，对天空中的每个指向产生 222 个天体辐射数据，这些数据经过前期预处理，会形成很多不同时间序列的数据，脉冲平均时间对于判断不同时间序列上的脉冲信号是否来自同一个天体有一定帮助。对于脉冲星信号，在多个子站可能同时被观测到，而对于 RFI 信号，往往只会在一个子站被观测到。

（9）信噪比分布曲线超额峰度（KurtSigma）：形成一个弥散脉冲组（DPG）的所有信号的信噪比分布曲线的峰度值减去正态分布的峰度，即图 1 前两个子图中曲线的峰度减去正态分布时的峰度，正态分布的峰度系数为 3。

（10）窗口宽度分布曲线超额峰度（KurtDuration）：形成一个弥散脉冲组（DPG）的每个事件在峰值检测时所用的窗口宽度值的分布曲线的峰度值减去正态分布的峰度，即图 1 中，后两个子图分布曲线的峰度值减去正态分布的峰度。

3. 方法

3.1 包裹式特征选择

特征选择的目的在于去除与当前学习任务无关和冗余的特征，降低学习任务的难度，促进对特征和问题的理解。其关键是建立一种评价标准来区分哪些特征组合有助于识别。为了增强特征与模型之间的相关性，提升模型性能，进行识别前，本文采用包裹法进行特征选择。

包裹式特征选择方法与后续任务选用的分类学习器直接相关，以学习器的性能作为特征子集的评估准则，即包裹式特征选择方法直接针对给定学习器进行优化（图 2）。因此，包裹式特征选择方法决策出的特征子集是最易与当前选用的分类器契合的。

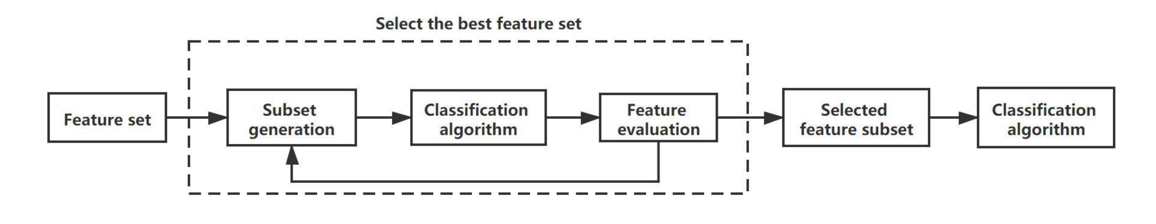


图 2　包裹式特征选择

Fig. 2 Wrapped method for feature selection

本文采用递归特征消除（Recursive feature elimination, RFE）的方法来实现包裹式的特征选择。分类器在给定的特征集合上进行训练，再从当前的特征集合中移除最不重要的特征，在新的特征集合上继续训练。不断重复递归这个过程，直到最终达到所需要的特征数量为止就决策出了最优的特征子集。对于给定分类器，最后选出的特征子集包含的特征就是最重要

的特征。

3.2 XGBoost 分类学习器

XGBoost 是一种集成学习算法，在决策树的基础上采用集成策略。XGBoost 包含一个迭代残差树的集合，利用梯度提升算法不断减小已生成的决策树的损失，每一棵树都在学习其前面所有树的残差，将每棵树预测的结果值相加作为样本的最终预测结果。

XGBoost 利用向前分布算法，学习到包含 K 棵树的加法模型：

$$\hat{y}_i = \sum_{t=1}^{K} f_t(x_i), f_t \in F \tag{1}$$

其中 K 为树的总棵树， f_t 表示第 t 棵树， x_i 表示输入样本， $\hat{y}_i$ 表示预测结果， $f_t(x_i)$ 表示第 t 棵树的预测结果，F 表示决策树组成的函数空间。

为了求解整个决策树的函数空间，需要不断优化目标函数，XGBoost 的整体目标函数可表示为[21]：

$$Obj(t) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{K} \Omega(f_t) \tag{2}$$

其中 $l(y_i, \hat{y}_i)$ 为损失函数，表示预测值 $\hat{y}_i$ 与目标值 y_i 之间的差值， $\Omega(f_t)$ 为第 t 棵树的正则项，用来约束决策树的复杂度，决策树的复杂度越高，正则项越大。

首先，通过贪心算法寻找局部最优解：

$$\hat{y}_i^{(t)} = \sum_{j=1}^{t} f_j(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{3}$$

$\hat{y}_i^{(t)}$ 表示第 t 次迭代时的第 i 棵树的预测结果，每次迭代寻找能使损失函数最大程度降低的 f_t。此时，目标函数可改写为：

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{4}$$

其次，对目标函数采用二阶泰勒近似展开得到：

$$Obj^{(t)} = \sum_{i=1}^{n} (l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)) + \Omega(f_t) \tag{5}$$

其中 g_i 和 h_i 分别表示误差函数的一阶导数和二阶导数：

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}),$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)})$$

移除对第 t 轮迭代来说为常数项的 $l(y_i, \hat{y}_i^{(t-1)})$ 得到：

$$Obj^{(t)} = \sum_{i=1}^{n}(g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)) + \Omega(f_t) \tag{6}$$

XGboost 中正则项用来衡量树的复杂度：

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \tag{7}$$

其中 T 表示每棵树的节点数量，ω 为每棵树的叶子节点的输出分数，γ, λ 为常数。可以进一步地将目标函数表示为：

$$Obj^{(t)} = \sum_{i=1}^{n}(g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)) + \gamma T + \frac{1}{2}\lambda \sum_{j}^{T} \omega_j^2 \tag{8}$$

将输入的 x_i 映射到叶子节点上，则有：$f_t(x_i) = \omega_{q(x_i)}, \omega \in R^T, q: R^d \to \{1,2,\ldots,T\}$，并定义每个叶子节点 j 上的样本集合为 $I_j = \{i \mid q(x_i) = j\}$

此时，目标函数可以表示为：

$$Obj^{(t)} = \sum_{j=1}^{T}(G_j \omega_j + \frac{1}{2}(H_j + \lambda)\omega_j^2) + \gamma T \tag{9}$$

其中 $G_j = \sum_{i \in I_j} g_i$，$H_j = \sum_{i \in I_j} h_i$

最后，对目标函数进行优化，计算第 t 轮时使目标函数最小的叶节点的输出分数 ω，直接对 ω 进行求导，使得导数为 0，得到：

$$\omega_j = -\frac{G_j}{H_j + \lambda} \tag{10}$$

将公式(10)带入(9)中，得到最终优化的目标函数：

$$Obj^{(t)} = -\frac{1}{2}\sum_{j=1}^{T}(\frac{G_j^2}{H_j + \lambda}) + \gamma T \tag{11}$$

在选择特征属性进行节点分裂时，XGBoost 会利用贪心算法或近似贪心算法，遍历所有特征的划分点，分别计算对应的目标函数值的增益，选择最优的特征进行分裂。当新的分裂带来的增益小于设定的阈值或达到设定的最大深度时，停止树的生长。XGBoost 对代价函数进行了二阶泰勒展开，还引入了缩减、行抽样和列抽样等操作，具有良好的预防过拟合、较高的计算效率和泛化能力的特性。对于 XGBoost 的程序实现可以直接采用基于 Python 语言

的机器学习工具包 Scikit-learn。

3.3 特征筛选评价流程

基于上述的理论和方法，本文将包裹式特征选择和 XGBoost 算法相结合。根据输入的数据集，设定一个阈值，取在该阈值下的最佳特征子集，并将得到的特征子集输入 XGBoost 算法用于分类，得到结果。具体流程图如图 3 所示。

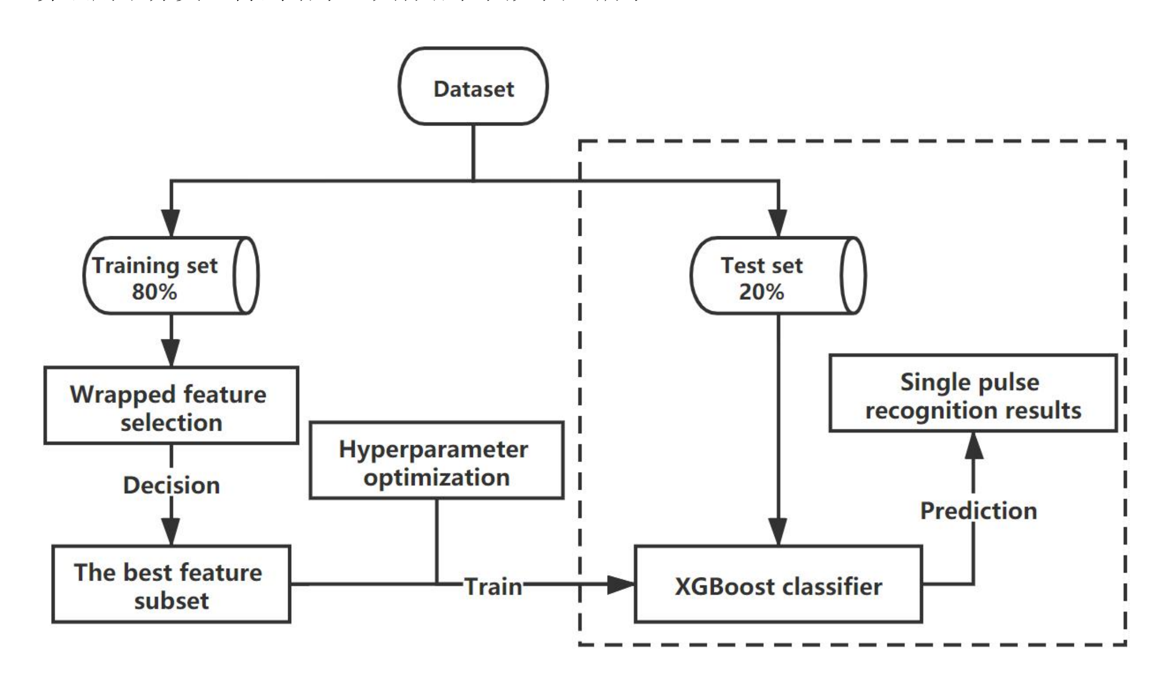


图 3 XGBoost 结合包裹式特征选择算法流程图

Fig. 3 Model flowchart of feture selection with XGBoost

为了分析当前方法的分类效果，我们利用混淆矩阵对模型的预测结果进行评价。本文数据集分为 RFI 弥散脉冲组（DPG）和单脉冲弥散脉冲组（DPG）。如果一个脉冲星的弥散脉冲组（DPG）被正确地识别为单脉冲信号，则我们称之为真阳性（True Positive，TP），若被错误地分类为 RFI 信号，则称之为伪阴性（False Negative，FN）。同样，如果一个 RFI 数据被错误地分类为单脉冲信号，则为假阳性（False Positive，FP），一个 RFI 数据被正确地分类为 RFI，则为真阴性（True Negative，TN）。表 1 为二分类情况下的混淆矩阵。

表 1 二分类混淆矩阵

Table 1 Confusion matrix for binary classification

Confusion Matrix		Target	
		Positive	Negative
Model	Positive	TP	FP
	Negative	FN	TN

在二分类问题中常用的评价指标有准确率（accuracy）、误分类率（error）、精确率（precision）、召回率（recall）和 F1 值（F1-score）[22]。其中，准确率表示正确分类的样本占总样本的比例，当数据集中存在各类别样本不平衡的情况时，分类器倾向将样本判断为来自比例较大的类别，出现准确率虚高的情况。因此，仅凭准确率并不能客观评价算法性能，还需要引入其他评价指标。精确率表示预测为脉冲星的样本中真正的脉冲星信号所占的比例。

召回率表示脉冲星信号被正确识别为脉冲星信号的比例。精确率和召回率两个指标存在此消彼长的问题，F1 值则综合了精确率和召回率的结果，可以调和平均两个指标，当前方法得到的 F1 值越高说明方法的性能整体上越理想。

4. 结果与讨论

实验所使用的数据集包含来自 47 个已知脉冲星的弥散脉冲组（DPG）18003 个，射频干扰弥散脉冲组（DPG）35063 个。具体操作步骤是将数据集随机划分 10 次，使用交叉验证方法进行模型的训练和评估，其中 80%用于训练，剩余的 20%用于验证。为了避免在数据分类中出现数据泄露，同时保证单脉冲样本和射频干扰样本尽量均衡，在实验中不是直接对数据集进行随机划分，而是分别将属于 47 个已知脉冲星的弥散脉冲组（DPG）和射频干扰弥散脉冲组（DPG）随机进行 10 次分组，每次分组将 80%，即 38 个已知脉冲星的弥散脉冲组（DPG）和射频干扰信号 80%的弥散脉冲组（DPG）用于训练，剩余 9 个已知脉冲星的弥散脉冲组（DPG）和射频干扰 20%的记录用于验证。

为了对比分析，本文除了采用 XGBoost 方法外，还对 GBDT、AdaBoost、Random forest 和 BP 神经网络模型（BPNN）进行了实验对比。为了使结果具有可对比行，对比前对每种方法都进行了调优，实验结果选用的都是调试出的最优参数，其中 BPNN 采用的是三层架构（输入层 10 个节点，隐藏层 56 个节点，输出 2 个节点），学习率为 0.0015，使用交叉熵损失函数和 Adam 优化器；GBDT 和 Random forest 的最大迭代次数是 100，最大深度是 20，学习率为 2；AdaBoost 的最大迭代次数是 100。表 2 显示了五种模型在该数据集上 10 次随机划分的平均实验结果。

表2 适用于不同模型的最优特征子集及平均实验结果

Table 2 Best feature sets and average results for different models

Model	Best feature subset	Error rate (Variance)	Precision (Variance)	Recall (Variance)	F1-Score (Variance)
XGBoost	Duration, DM, S/N, DM_E, Time_E, N_events, aDM, KurtDuration, KurtSigma	2.16% (0.0007)	97.57% (0.0012)	99.28% (0.0000)	98.41% (0.0003)
GBDT	Duration, DM, S/N, DM_E, Time_E, aDM, N_events, KurtDuration, KurtSigma	4.05% (0.0004)	96.14% (0.0010)	98.12% (0.0000)	97.09% (0.0002)
AdaBoost	Duration, DM, S/N, DM_E, Time_E, aDM, aTime, N_events, KurtDuration, KurtSigma	3.16% (0.0003)	96.53% (0.0007)	99.02% (0.0000)	97.74% (0.0002)
Random Forest	Duration, DM, S/N, DM_E, Time_E, aDM, N_events, KurtDuration, KurtSigma	3.95% (0.0013)	95.70% (0.0016)	99.18% (0.0000)	97.35% (0.0004)
BPNN	Duration, DM, S/N, DM_E, Time_E, aDM, aTime, N_events, KurtDuration, KurtSigma	3.98% (0.0001)	97.47% (0.0009)	98.11% (0.0001)	97.75% (0.0002)

包裹法特征选择的过程紧密结合选用的分类器，通过某种特征搜寻策略在模型上检验多种特征子集的分类性能，表 2 列出了不同分类器和其最优特征组合的分类评价结果。从表 2 可以看出，五种模型对脉冲星弥散脉冲组（DPG）和 RFI 弥散脉冲组（DPG）的分类结果都具有较高的精确率和召回率。特别是，XGBoost 的精确率、召回率与 F1-Score 是五种模

型中最高的，比其它模型平均高出 1 到 2 个百分点。从误分类率来看，GBDT 的误分类率最高，XGBoost 误分类率最低。综合几个指标的结果可以看出，XGBoost 在单脉冲信号弥散脉冲组（DPG）分类识别方面更有综合优势。

从特征应用上看，色散（DM）、信噪比（S/N）、窗口宽度（Duration）、色散范围（DM_E）、时间范围（Time_E）、事件个数（N_Events）、色散平均值（aDM）、窗口宽度分布曲线超额峰度（KurtDuration）、信噪比分布曲线超额峰度（KurtSigma）等九个特征参数被五个模型都看作是最优特征组合，AdaBoost 和 BPNN 模型进一步把脉冲平均时间（aTime）也选为最优特征。aTime 没有被其它三个模型选为最优特征，说明该特征对单脉冲识别的作用不是特别显著。

对 XGBoost 分类器，除了用于模型训练的特征之外，超参数也会在一定程度上会影响单脉冲识别的结果，其中树的最大深度、模型的学习速率是影响结果性能的主要参数。

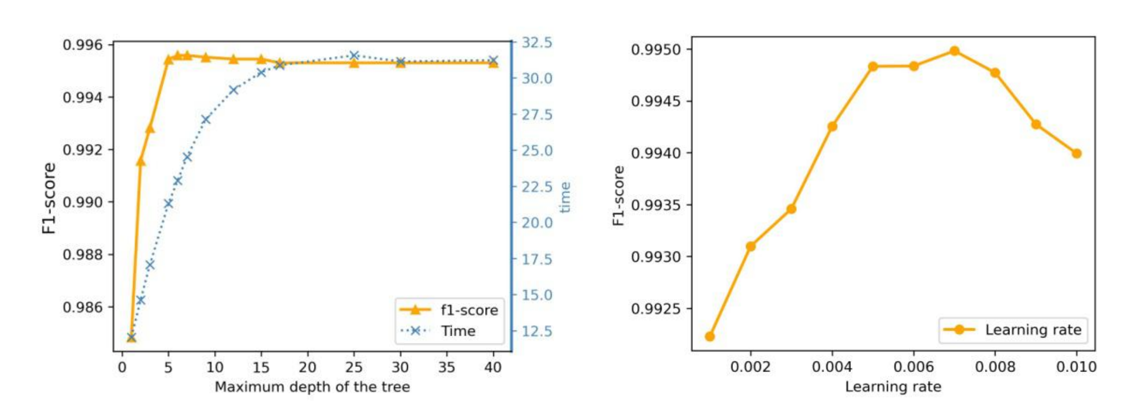


图 4 模型 F1 值与树的最大深度变化关系

Fig. 4 F1-score changes with the maximum depth of the tree

图 5 模型的 F1 值随学习率的变化趋势

Fig. 5 F1-score changes with model learning rate

图 4 所示为树的最大深度对模型的训练时间以及 F1 值的影响。当树的最大深度小于 25 时，训练模型花费的时间稳步上升，而后基本保持平稳不变；而模型的 F1 值随着树的最大深度变化呈现出先升后降，而后平稳的趋势。当树的最大深度取值为 6 时，XGBoost 能够在测试集上获得一个最高的 F1 值，且用时相对来说也较短。由此可见，在本文使用的数据集上，树的最大深度在 6 时能同时权衡训练模型的时间消耗和单脉冲分类任务的性能。图 5 展示了学习速率对 XGBoost 的性能影响。由图 5 可知，在学习率达到 0.007 的时候，XGBoost 能获得最好的分类性能。

特征数量同样也会影响模型对单脉冲识别的性能。在本文所使用的数据集上对 10 个特征使用 XGBoost 结合包裹式特征选择算法对特征重要程度进行分析。针对包裹式特征选择算法，我们通过设置不同阈值获得不同规模的最优特征子集，并比较分析基于这些特征子集的模型性能。表 3 展示了基于不同规模特征子集训练的 XGBoost 模型在单脉冲信号识别任务上的 F1 值。

表 3 XGBoost 模型基于不同规模的最优特征子集训练时分类的 F1 值

Table 3 F1-scores for different feature sets with XGBoost classification

Feature count	The best feature subset	F1-score
4	Duration, DM, S/N, DM_E	0.9706
5	Duration, DM, S/N, DM_E, aDM	0.9678
6	Duration, DM, S/N, DM_E, aDM, Time_E	0.9741
7	Duration, DM, S/N, DM_E, aDM, Time_E, N_events	0.9769
8	Duration, DM, S/N, DM_E, aDM, Time_E, N_events, KurtSigma	0.9838
9	Duration, DM, S/N, DM_E, aDM, Time_E, N_events, KurtSigma, KurtDuration	0.9866
10	Duration, DM, S/N, DM_E, aDM, Time_E, N_events, KurtSigma, KurtDuration, aTime	0.9845

结果表明，特征的数量也会影响脉冲信号分类的性能。虽然每个特征对模型的影响不同，但输入特征的数量和组合也是影响模型性能的关键因素。可以看出，输入不同数量的特征会得到不同的结果。当特征数量为 9，特征为 Duration, DM, S/N, DM_E, aDM, Time_E, N_events, KurtSigma, KurtDuration 时，可以获得最高的 F1 值。

5. 结论

近年来，随着周期性脉冲信号探测方法的成熟，对单脉冲信号的识别成为脉冲星研究的一个重要领域。由于单脉冲信号可以提取的特征相对较少，机器学习方法成为最主要的方法。设计关键特征，并寻找最优的机器学习算法是当前脉冲星信号识别的关键任务。

本文在前人研究的基础上，将 XGBoost 分类器与包裹式特征选择相结合，以 LOTAAS 数据集为例，与 AdaBoost、GBDT、Random Forest 和 BP 神经网络等模型进行了实验对比。研究结果表明，XGBoost 在单脉冲识别方面误分类率更低，精确率、召回率与 F1-值更高，是进行单脉冲信号识别提取的优秀方法。本文在实验设计中，将 47 个已知脉冲星和射频干扰信号分别随机进行 10 次分组，有效避免了数据集划分造成数据泄露的影响。如果直接将来自 47 个脉冲星的 18003 个弥散脉冲组（DPG）进行分组训练和测试，得到精确度将高达 99.79%，F1-score 高达 99.76%。可见训练集和测试集的划分方法对识别结果具有重要影响。

从特征选择上看，本文的实验结果表明色散、信噪比、窗口宽度、色散范围、时间范围、事件个数、色散平均值、窗口宽度分布曲线超额峰度、信噪比分布曲线超额峰度等九个特征被最多模型选择，具有良好的判别力。

对单脉冲信号进行标注建立训练数据集是一件费力耗时的工作，需要长期的积累。本文虽然是以 LOTAAS 数据集作为研究对象，其研究结果和方法可以为我国开展以 FAST 探测信号为主的单脉冲信号研究与应用提供参考。目前对我国 FAST 数据的挖掘和应用正在大力推进中，并已成功探测到属于单脉冲的快速射电暴[23][24]。另外，随着对单脉冲信号特征的持续分析和挖掘，新的研究方法也将不断提出和改进。

参考文献

[1] Antony Hewish. Pulsars as Physics Laboratories[J],Interdisciplinary Science Reviews, 1994, 19:1,70-74.

[2] Cordes J M; McLaughlin, M A Searches for Fast Radio Transients[J], The Astrophysical Journal, 2003, 596:2, 1142-1154.

[3] Thomas Ryan Devine, Katerina Goseva-Popstojanova, Maura McLaughlin, Detection of dispersed radio pulses: a machine learning approach to candidate identification and classification[J], Monthly Notices of the Royal Astronomical Society, 2016, 459:2, 1519–1532.

[4] Patel C, Agarwal D, Bhardwaj M, Boyce M M, Brazier A, & Chatterjee S, et al. Palfa single-pulse pipeline: new pulsars, rotating radio transients, and a candidate fast radio burst[J]. Astrophysical Journal, 2018, 869(2).

[5] McLaughlin M, Lyne A, Lorimer D, et al. Transient radio bursts from rotating neutron stars[J]. Nature, 2006, 439, 817–820.

[6] Lorimer D R, Bailes M, McLaughlin M A, Narkevic D J & Crawford F A. bright millisecond radio burst of extragalactic origin[J]. Science, 2007, 318, 777–780.

[7] Deneva J S , Cordes J M , Mclaughlin M A , et al. Arecibo Pulsar Survey Using ALFA: Probing Radio Pulsar Intermittency and Transients[J]. Astrophysical Journal, 2009, 703(2):2259-2274.

[8] Karako-Argaman C, et al. Discovery and Follow-up of Rotating Radio Transients with the Green Bank and LOFAR Telescopes[J]. The Astrophysical Journal, 2015, 809(1):67.

[9] Ryan D T , Katerina G P , Maura M L . Detection of Dispersed Radio Pulses: A machine learning approach to candidate identification and classification[J]. Monthly Notices of the Royal Astronomical Society, 2016(2):stw655.

[10] 王元超，郑建华，潘之辰，李明涛. 脉冲星候选样本分类方法综述[J]，深空探测学报，2018, 5:3, 203-211.

[11] McFadden R, Karastergiou A, Roberts S. Machine learning for pulsar detection[J]. Proceedings of the International Astronomical Union, 2017, 13(S337): 372-373.

[12] Eatough R P , Molkenthin N , Kramer M , et al. Selection of radio pulsar candidates using artificial neural networks[J]. Monthly Notices of the Royal Astronomical Society, 2010, 407(4).

[13] Bates S D , Bailes M , Barsdell B R , et al. The High Time Resolution Universe Pulsar Survey – VI. An artificial neural network and timing of 75 pulsars[J]. Monthly Notices of the Royal Astronomical Society, 2012, 427(2):1052-1065.

[14] Morello V , Barr E D , Bailes M , et al. SPINN: a straightforward machine learning solution to the pulsar candidate selection problem[J]. Monthly Notices of the Royal Astronomical Society, 2014, 443(2).

[15] Zhu W W , Berndsen A , Madsen E C , et al. Searching for Pulsars Using Image Pattern Recognition[J]. The Astrophysical Journal, 2014, 781(2):117.

[16] Michilli D , Hessels J , Lyon R J , et al. Single-pulse classifier for the LOFAR Tied-Array All-sky Survey[J]. Monthly Notices of the Royal Astronomical Society, 2018(3):3.

[17] Chen T , Guestrin C . XGBoost:A scalable tree boosting system[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 16. ACM, 2016:785-794.

[18] Coenen T, van Leeuwen J, Hessels J W T, et al. The LOFAR pilot surveys for pulsars and fast radio transients[J]. Astronomy & Astrophysics, 2014, 570.

[19] Sanidas S, Cooper S, Bassa C G, Hessels J W. T, Kondratiev V I, Michilli D, et al. The LOFAR Tied-Array All-Sky Survey (LOTAAS): Survey overview and initial pulsar discoveries[J]. Astronomy & Astrophysics, 2019, 626.

[20] Lorimer D R, Kramer M. Handbook of Pulsar Astronomy[DB/OL], 2004.

[21] 李航. 统计学习方法[M]. 清华大学出版社, 2012.

[22] Suthaharan S. Machine learning models and algorithms for big data classification[M]. New York: Springer, 2016.

[23] Weiwei Zhu, Di Li, Rui Luo, et al. A Fast Radio Burst discovered in FAST drift scan survey[J]. ApJL, 2020, DOI: 10.3847/2041-8213/ab8e46

[24] Chen-Hui Niu, Di Li, Rui Luo, et al. CRAFTS for Fast Radio Bursts Extending the dispersion-fluence relation with new FRBs detected by FAST[J]. ApJL, 2021, DOI: 10.3847/2041-8213/abe7f0

Research on recognition method of single-pulse signals based on XGBoost

Yu Ling, Jinqu Zhang, Xiangru Li, Hui Li
School of Computer Science, South China Normal University, Guangzhou, 510631, China

Abstract:
With the construction of large-scale radio telescopes, detecting pulsars from large-scale pulse signals has become an important task of space exploration. Machine learning algorithms are favored in single-pulse data analysis due to their data-driven advantages. However, algorithms used in pulsar searching cannot guarantee that their results are global optimal solutions. In this paper, eXtreme Gradient Boosting (XGBoost) method is studied in single pulse classification with the data from the LOFAR Tied-Array All-Sky Survey (LOTAAS). The LOTAAS is is an ongoing survey of the Northern sky for pulsars and transients with LOFAR using a digital aperture array. As of January 2019, the LOTAAS survey has discovered and confirmed 73 radio pulsars, which demonstrates its ability to find new pulsars. A fully labeled data set used for training and validation of the machine model is necessary. However, faced with massive amounts of astronomical observation data, it's time-consuming and laborious work to labeling data with manual inspection. In this study, we directly use the well-prepared data in the work of Michilli et al. (2018) for saving the labor of repetitive processing of data. In order to verify the performance of XGBoost method, this paper compares the algorithm with other four machine learning models,. The results show that XGBoost combined with wrapped feature selection method has more advantages in single pulse recognition, with the lowest misclassification rate and the highest accuracy, and F1 score. This study has important implications for pulsar monitoring and can provide a reference for the research of single pulse search based on Five-hundred-meter Aperture Spherical radio Telescope (FAST) signals in China.

Keywords: Single-pulse; XGBoost; Feature selection; Wrapping method